International Journal of Combinatorial Optimization Problems and Informatics, 12(3), Sep-Dec 2021, 87-97. ISSN: 2007-1558.

www.editada.org

New approach to feature extraction in authorship attribution

Omar González Brito[1], *José Luis Tapia Fabela*[1], *Silvia Salas Hernández*[2]
[1] Universidad Autónoma del Estado de México, Unidad Académica Profesional Tianguistenco, Paraje el Tejocote s/n, San Pedro Tlaltizapán, 52640, Santiago Tianguistenco, Méx.
[2] Universidad Autónoma del Estado de México, Centro Universitario Atlacomulco, km 60, carretera Toluca-Atlacomulco, 50450, Atlacomulco, Méx.
gonzalezbritoomar@gmail.com, joseluis.fabela@gmail.com, salashernandezsilvia@gmail.com

Abstract. Conventionally, the authorship attribution has been carried out through text classification strategies; at the same time, the feature extraction process has been carried out under two main approaches: instance-based and profile-based. However, these approaches generate a high feature dimensionality that can impair the classification performance; furthermore, it is required to consider a feature selection method. This work proposes an approach that does not depend on the total number of documents for the feature extraction. In our research, we shown that the features that describe an author's writing style can be contained in a single document. To carry out this experimentation, we worked with three corpuses (C10, C50, and PAN12), which were selected based on the literature review. According to the results obtained, with a classification accuracy of 79.68%, it is concluded that the proposed approach presents superior results to the state of the art using imbalanced samples. In addition, the approach is robust when it is evaluated in different contexts. Finally, from the experimentation, it is determined that in approximately 500 words without repeating the writing style of an author is contained.
Keywords: Authorship attribution, feature extraction, text classification, supervised learning, authorship approaches.

Article Info
Received Jan 18, 2021
Accepted April 17, 2021

1 Introduction

At present, authorship analysis has become a major problem in many areas, including information retrieval, computer linguistics, and forensic linguistics. The main conflicts related to authorship occur in plagiarism in student essays, forensic cases, cyberbullying, fraud detection, social media, messages, and emails [1,2,3,4,5,6].
The computational problem of authorship attribution has been analyzed in the state of the art mainly through the automatic classification of texts [7,8]. The author of a document is identified by carrying out the analysis of textual features (lexical, syntactic, semantic, and morphological) which allows us to determine whether or not a document belongs to an author among a group of candidates [9,10,11,12].

Although the methodology of classification of texts consists of the following stages: 1.- Data acquisition, 2.- Data analysis and labeling, 3.- Feature construction and weighting, 4.- Feature selection and projection, 5.- Training of a classification model, 6.- Solution evaluation [13]. The present research focuses on the stage of data analysis and labeling; at this stage, the features are extracted to be able to compare them with the document in question and determine their authorship. In the state of the art, feature extraction is mainly done under two approaches: profile-based and instance-based [1], [14]. The profile-based approach extracts the features from the file resulting from the concatenation of all training documents for an author [1,14]. In the instance-based approach, the feature extraction process is performed by each author's document [1,14,15,16,17,18]. However, these two approaches produce high dimensionality of features and they are dependent on the size of the training sample. If you have few documents from an author, the classifier's performance decreases dramatically [19].

This research proposes a new approach to the task of authorship attribution for cases where few documents from an author are available. This is due to the fact that on many occasions when state-of-the-art methods are used, the total

number of documents required to be able to identify the authors is not available. The contribution of our approach is the extraction of features from a single random document for each author of the training set. The question we want to answer in this work is: does a single document have the necessary characteristics to determine an author's writing style? To answer it, lexical features (n-grams and bag of words) with a Boolean representation are used, because they are considered to contain the writing style of an author and they are the ones with which the best results have been obtained in the representation of the texts [20], [21]. Moreover, the supervised learning methods used are Support Vector Machine, multinomial logistic regression, Naive Bayes, decision trees, and random forests. Finally, the evaluation was performed using three corpuses C10, C50, and PAN12-I, different number of authors, contexts and sample sizes.

2 State of the art

One of the main problems in the authorship attribution task is determining the size of the data set from which it is possible to distinguish the authorship of a text [14]. Maciej-Eder [22] proposes the analysis of the length of the sample from a corpus formed by different literary genres, he takes a long text and segments it into shorter texts; These are analyzed to determine the authorship of the original texts. The size of the analyzed texts was 500, 600, 700, ... and 20,000 words. Being the ranges from 2500 to 5000 where he obtained better results. Besides, Gómez-Adorno [14], through the use of syntactic graphs, extracts the necessary patterns to attribute authorship through the cosine similarity measure. The method of Gómez-Adorno [14] does not use classification algorithms and can be implemented in the tasks verification and authorship attribution, using a reduced set of labeled data, the results obtained to determine authorship is good.

From the works of Maciej-Eder [22] and Gómez-Adorno [14], the premise arises that there are enough features to determine authorship in small sets of documents. Furthermore, analyzing the corpuses C10 and C50 it is obtained that the number of words without repeating per document is between 500 and 506 words. As a result of the analysis of the previously described works, our idea arises to implement an approach for the feature extraction that does not require the entire data set; and with this, determine if in a single document there are the necessary features to carry out the attribution of authorship.

In order to evaluate the proposed approach, the text classification method is used, since it is one of the most used in the task [13]. The contributions using this method have been given mainly in the stages of feature construction and weighting [23], [17], feature selection and projection [19] and training a classification model [4], [24], [25]. Hence the importance of our work since it focuses mainly on the data analysis and labeling stage that has been little studied.

From the analysis of the state of the art, the following attributes are considered. The proposed approach was implemented for the data analysis and labeling stage. For the construction and weighting of features, we used Boolean weighting through vector representation [17]. For model training we used Support Vector Machine, logistic regression, multinomial Naive Bayes, decision tree and random forest [4], [24]; Finally, the approach is evaluated through the accuracy metric using a different number of authors, contexts, and balanced and unbalanced samples [26], [27].

3 Materials and methods

Based on the literature review there are two approaches used in the authorship attribution task; profile and instances, during the process of extracting features in these approaches a high dimensionality is generated, where relevant, redundant, and irrelevant features are present. Redundant and irrelevant features impair the performance of the classifier. The present research proposes a new approach to feature extraction, which consists of the random selection of a document from the training set for feature extraction.

To evaluate the approach, a text classification method is implemented that consists of the following steps: 1.- Data Acquisition, 2.- Data analysis and labeling, 3.- Feature construction and weighting, 4.- Feature selection and projection, 5.- Training of a classification model, 6.- Solution evaluation [13]. In the present investigation, the stage of construction of features and weighting is carried out under the proposed approach while the stage of selection and projection of features is omitted. The classification method for the assessment of the approach consists of the following steps.

3.1 Data acquisition

In 2009, the first international plagiarism detection competition was held as part of the PAN workshop (CLEF, Conference, and Labs of the Evaluation Forum) its objective has been to encourage the development of automatic tools for the detection of plagiarism and currently also the identification of authorship and other abusive uses of social software [28]. In the web page of PAN (https://pan.webis.de/) are the corpuses C10, C50, and PAN12-I, this corpus has been used for the task of authorship attribution in different works of the literature. The access to these corpuses is public.

3.2 Data analysis and labeling

Feature extraction is done using the proposed approach. This consists of the random selection of a document from the training set for each author. They are then represented in the vector model, which will be used by the text classification method to generate the attribution model as shown in Fig. 1. Through experimentation, it was verified that the necessary features to determine the authorship exist in a single document.

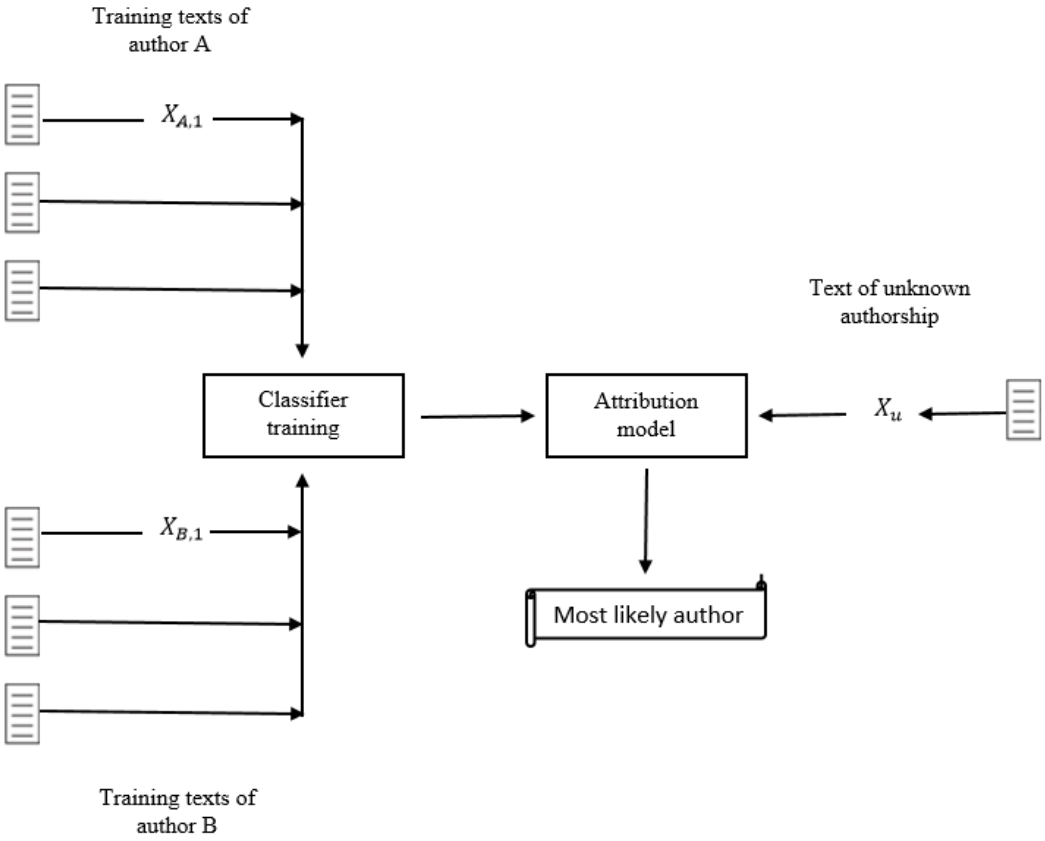

Fig. 1 New approach to feature extraction.

For the labeling of data, the models of representation of bag of words and n-grams were used; because they contain the style of writing of an author and are with those that have obtained better results in the representation of texts [20], [21]. For the word bag representation, the following special characters were removed, punctuation, admiration, and question marks. The trigrams model does not perform any pre-processing.

3.3 Feature construction and weighting

Boolean or binary weighting consists of assigning a value to a term within a document. The value assigned to the term reflects the importance of the term. A value of one is assigned when the term is present otherwise a value of zero is assigned. Through this weighting, you can see the importance of each of the terms. The above is represented through equation 1, where: tj is the frequency of the term j that has the sentence pi [29].

$$p_{i=(t_j)}=\begin{cases}1, & if\ it\ exists\\ 0, & otherwise\end{cases} \quad (1)$$

Table 1 shows the vector representation used in this research; where, the row represents the document of each author and the columns the terms. subsequently, the table is filled with the Boolean weighting, where 1 means the presence of the term within the author's document, otherwise, a 0 is placed. This process is done until all the authors' documents are finished.

Table 1. Vector representation with Boolean weighting.

Document/author	**Term 1**	**Term2**	**Term 3**	**Term …**
Document 1 author1	0	0	1	1
.	1	1	0	0
.	0	1	1	0
.	1	1	0	1
Document n autor 1	0	0	1	0
Document 1 author 2	1	1	0	1
.	0	0	0	0
.	1	1	1	0
.	0	0	1	1
Document n author 2	1	1	0	1
Document 1 author 3	0	0	1	0
.	0	1	1	0
.	0	0	0	0
.	1	0	1	1
Document n author 3	1	0	1	1
…				

3.4.-Training of the model

The model was built from supervised learning. It was implemented with a support vector machine (SVM), multinomial logistic regression, Naive Bayes, decision tree and random forests. The SVM parameters were a linear kernel, the parameter C equal to one, a classification of one against all was used. The SVM was trained with the features of each of the authors. Once the model was generated, the evaluation was carried out.

3.5 Evaluation of the solution

The metric used for the evaluation was accuracy as shown in equation 2 [17]. This consists of the percentage of instances that are correctly classified, defined in terms of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

4 Experimentation and results

To verify the proposed approach, 7 experiments were carried out. These were validated using the ten-fold cross-validation technique, using the text classification method described in section 3, materials and method. The models used for the representation of characteristics are a bag of words and trigrams. The learning methods used are vector support machine, logistic regression, Naive Bayes, decision trees, and random forests. The accuracy metric was used for the evaluation of the results. The experimentation was carried out with the corpuses C10, C50, and PAN12 [17], [18], [19]. Table 2 shows the structure of the corpuses.

Table 2. Structure of the corpuses used.

Corpus	Number of documents	Number of authors	Training and validation documents	The average number of words per document	Category
C50	5000	50	50 / 50	500	Documents belonging to the category CCTA industry news
C10	1000	10	50 / 50	507	
PAN12-I	42	14	2 /1	3345	Literature and novels

The experiments are described below:

Experiment 1: Through Fig. 2 the robustness of the proposed approach in the C10 corpus is observed by using a word bag representation with 50 documents versus 1 document for the feature extraction. Different supervised learning models were used to analyze their performance, observing that for authorship attribution the significance of using 50 documents or 1 document for feature extraction does not impair the performance of the classifier.

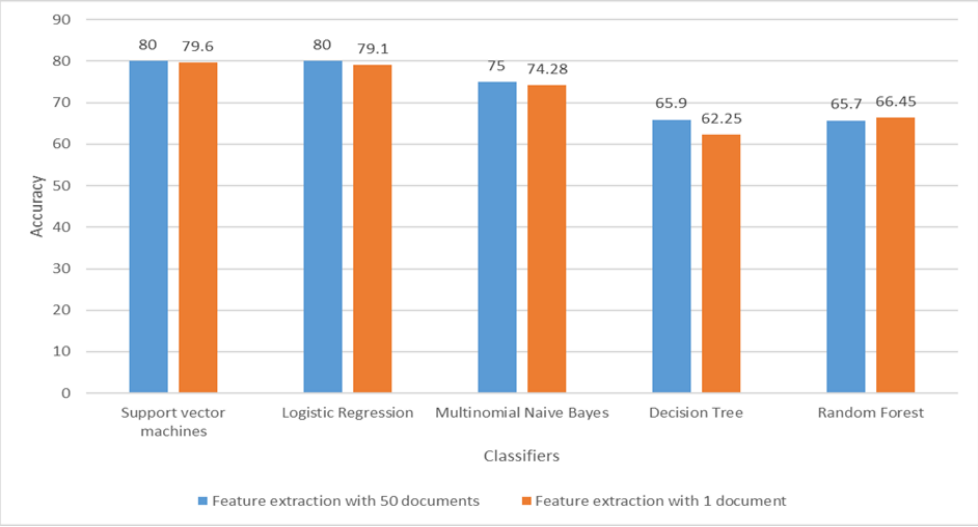

Fig. 2. Evaluation of the feature extraction process with different classifiers with the word bag model for the C10 corpus.

Experiment 2: Using a representation of trigrams with 50 documents versus 1 document, it is observed in Fig. 3 again that the proposed approach shows robustness in the C50 corpus using different classifiers. The results obtained for the case of Naive Bayes, decision trees, and random forests are higher than for the case of 50 documents. This proves that extracting features from a single document performs better than extracting features from the entire document set.

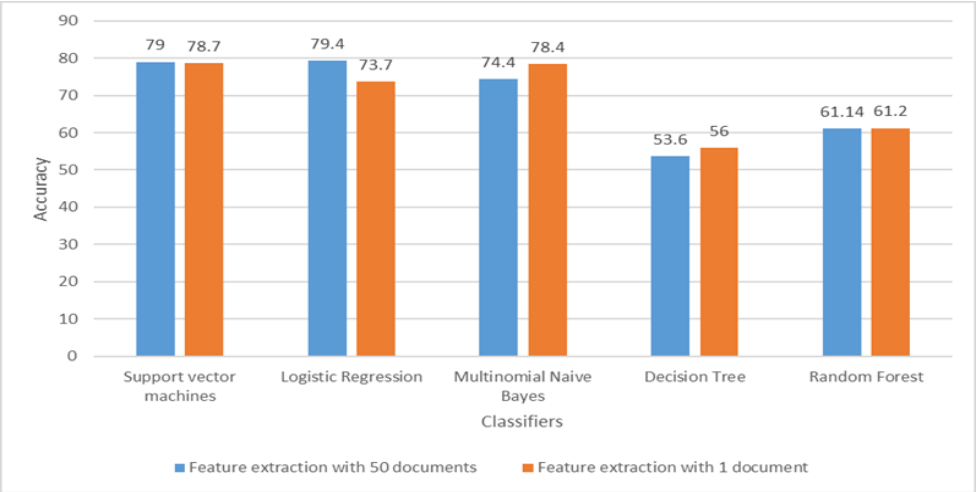

Fig. 3. Evaluation of the feature extraction process with different classifiers with the trigrams model for the C50 corpus.

Experiment 3: In this experiment, the importance of the number of documents in the extraction of the characteristics were analyzed. Fig. 4 shows that the best results are obtained when three to six documents are used. When three documents are used, there is an accuracy of 82% which is higher than the 80.8% reached by [19]. It can also be observed that the increase in attributes for the extraction of features does not show a considerable improvement in the accuracy of the classifier. It must be considered that characteristic selection methods, which eliminated redundant and irrelevant features, were not used.

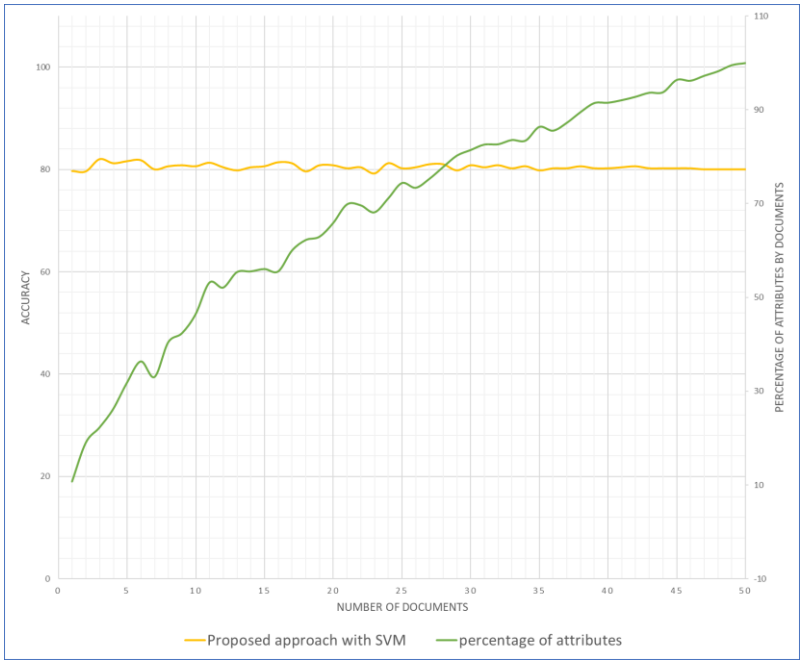

Fig. 4. Analysis of the number of documents for author.

Experiment 4: The objective of this experiment was to verify the robustness of the proposed method using the C10 corpus. In Fig. 5. the results of the proposed method are compared with the tensor space model [19], the Doc2vec method [17], and with patterns extracted methods presented in [14]. The proposed method reaches an accuracy of 79.68%, which represents a difference of 1.12% concerning the model of space of tensors presented in [19], from above we can conclude that the proposed method has a good performance in balanced samples.

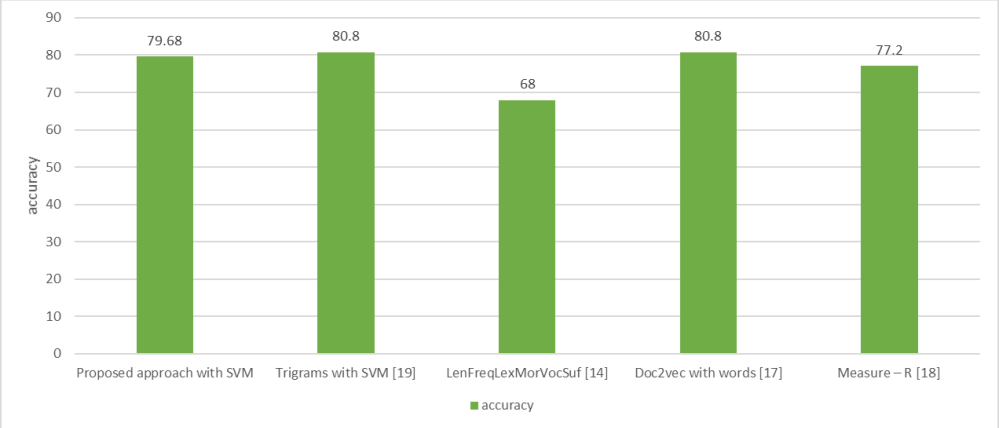

Fig. 5. Performance of the method proposed versus the state of the art, corpus C10.

Experiment 5: This experiment is done in the same way as the first one but with the corpus C50. The C50 corpus contains 50 authors each with 50 training documents and 50 validation documents unlike the C10 containing 10 authors. The objective of this experiment is to know if the method obtains a good accuracy in the classification with fifty authors. The accuracy obtained is 72.04, a competitive result concerning the state of the art as can be seen in Fig. 6.

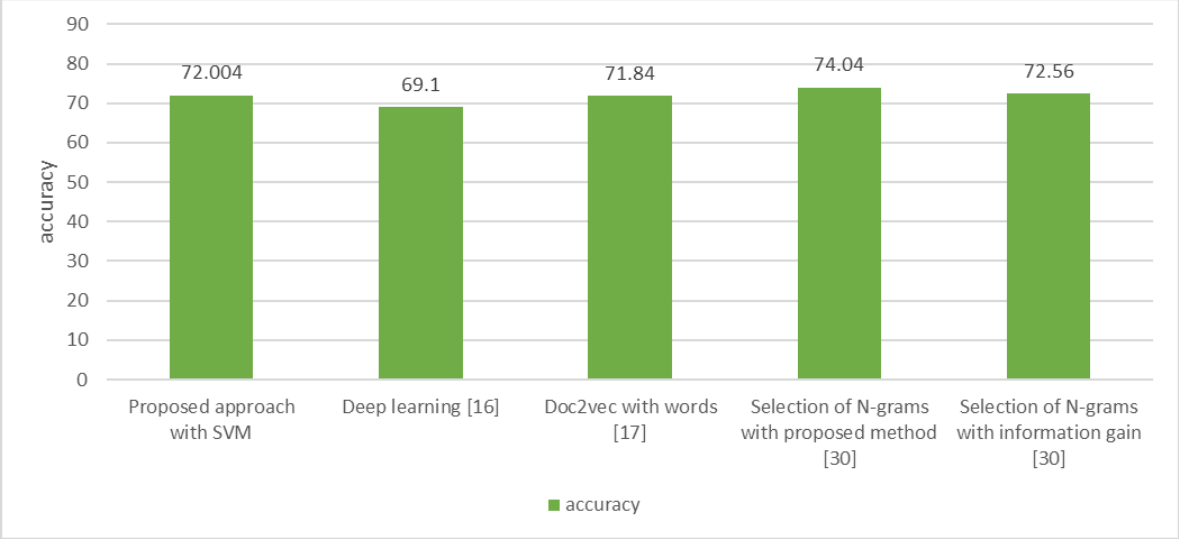

Fig. 6. Performance of the proposed method versus the state of the art, corpus C50.

Experiment 6: The proposed approach was analyzed under two different contexts: emails and literary works. The results were compared with the work of [18], where they reproduce the most referenced methods of the state of the art for authorship attribution. The proposed method obtains better results than some of the methods evaluated by [18]. It can be seen in Fig. 7 that the proposed method remains stable in the two corpuses, obtaining results greater than 70%, in comparison to those evaluated by [18]. In the state of the art, it is common that the methods obtain good results with one corpus and low results with the other.

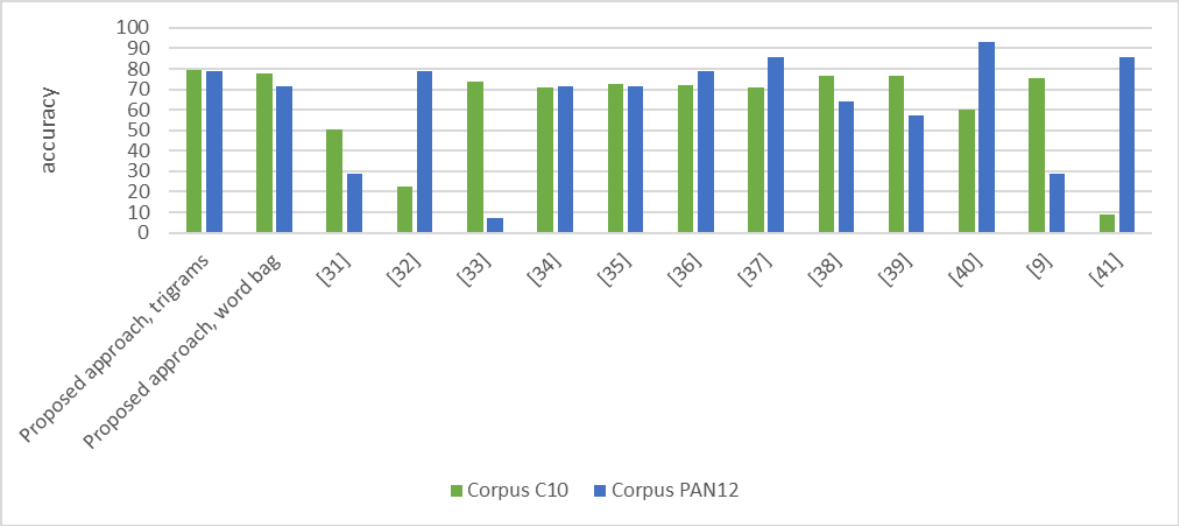

Fig. 7. Evaluation of the proposed method under different contexts.

Experiment 7: The analysis of balanced and unbalanced samples is treated in the state of the art using the C10 corpus. In this experiment, the proposed approach is analyzed considering as features the trigrams and bag of words. As can be

seen in Fig. 8, the proposed method obtains better results in imbalanced samples than the one proposed by [19]; even in balanced samples, competitive results are obtained. From above we can conclude that the proposed method is also robust with balanced and imbalanced samples.

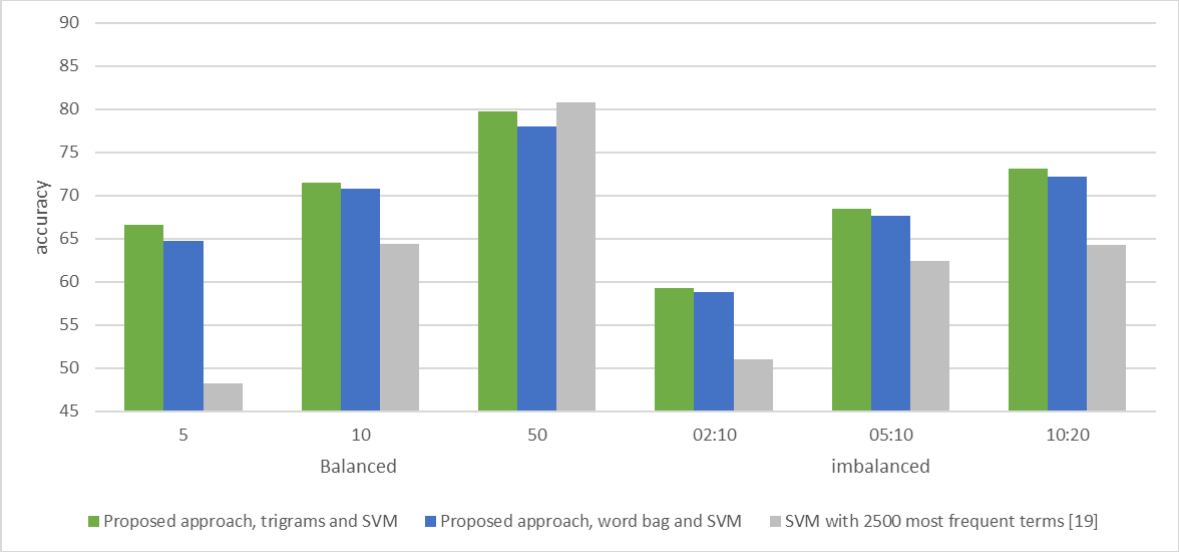

Fig. 8. Analysis of balanced and unbalanced samples in Corpus C10.

4 Conclusions

Through experiments 1, 2, 4, and 5 it is concluded that an author's writing style is contained in a single document. When extracting the characteristics of 50 documents versus 1 document, it is determined that the difference is less than 1% of accuracy, as can be seen in Fig. 2 and Fig. 3. Comparing the proposal with the state of the art of C10 corpus, it obtains competitive results as you can see in Fig. 5 the difference is 1.12% with respect to the best state-of-the-art method. Fig. 6 corresponds to an experiment with the C50 corpus where there is a difference of just 2.04%.
In experiment 6, the analysis of the proposed approach is carried out under different contexts using the C10 and PAN12 corpus. In Fig. 7 it is observed that the proposed approach is consistent under different contexts showing robustness. The analysis of the methods presented by [18] shows that some methods present variability in their results when working under different contexts. An example of this is the method presented in [33] where the result with the C10 corpus is 78% accuracy and with the PAN12 corpus it is 7.1% another case is the method presented in [41] where the result with the C10 corpus is 9% accuracy and with the PAN12 corpus the accuracy is 85.7%.

In experiment 7 the analysis of balanced and imbalanced samples was carried out; this analysis can be transferred to real-life where in some cases there is not the same number of documents for each author. As can be seen in Fig. 8 for the analysis of unbalanced samples, our approach obtains better results than those presented in [19], where in the 2:10 sample it is exceeded by 8.3%. For the case of the 5:10 sample, the percentage with which it is exceeded is 6% and for the 10:20 sample is 9.92%. For the balanced samples, the approach exceeds two of the three samples of sizes 5 and 10 with a difference of 18.32% and 7.04% respectively. However, in the sample of size 50, the method proposed by [19] beats our method by 1.12% as observed in Fig. 8. Throughout these results it is determined that the proposed approach is robust when analyzing unbalanced samples, exceeding the state of the art as seen in Fig. 8.

One area of opportunity that arises from this research is the development of methods that can determine authorship from a single document by implementing similarity metrics.

References

1. Stamatatos. E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology., 60(3),538-556,
2. Lambers, M., & Veenman, C. J. (2009). Forensic authorship attribution using compression distances to prototypes. In International Workshop on Computational Forensics, Springer, Berlin, Heidelberg, 13-24.
3. Pillay, S. R., & Solorio, T. (2010). Authorship attribution of web forum posts. In 2010 eCrime Researchers Summit, 1-7
4. Ramnial, H., Panchoo, S., & Pudaruth, S. (2016). Authorship attribution using stylometry and machine learning techniques. In Intelligent Systems Technologies and Applications, 113-125
5. Zhang, S. (2016). Authorship attribution and feature testing for short Chinese emails. International Journal of Speech, Language & the Law, 23(1).
6. Shrestha, P., Sierra, S., González, F. A., Montes, M., Rosso, P., & Solorio, T. (2017). Convolutional neural networks for authorship attribution of short texts. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2, 669-674
7. Xue. B, Zhang. (2015). A comprehensive comparison on evolutionary feature selection approaches to classification, International Journal of Computational Intelligence and Applications, 14(02), 1550008.
8. Qian. C, He. T, Zhang. R. (2017). Deep Learning Based Authorship Identification
9. Escalante, H. J., Solorio, T., Montes-y Gómez, M. (2011). Local histograms of character n-grams for authorship attribution. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 288-298.
10. Kern, R., Seifert, C., Zechner, M., & Granitzer, M. (2011). Vote/veto meta-classifier for authorship identification. In CLEF 2011: Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation (Lab and Workshop Notebook Papers), Amsterdam, The Netherlands.
11. Solorio, T., Montes, M., & Pillay, S. (2011). Authorship Identification with Modality Specific Meta Features. Notebook for PAN at CLEF 2011.
12. Castro, D. Adame, Y. Peláez, M. Muñoz, R. (2015). Authorship verification, average similarity analysis. International Conference Recent Advances in Natural Language Processing, Vol.902015, pp.84-902015, (2015)
13. Mirończuk. M, Protasiewicz. J, "A recent overview of the state-of-the-art elements of text classification", Expert Systems with Applications, Vol.106, pp.36-54, (2018)
14. Gómez-Adorno, H., Sidorov, G., Pinto, D., Vilariño, D., & Gelbukh, A. (2016). Automatic authorship detection using textual patterns extracted from integrated syntactic graphs. Sensors, 16(9), 1374.
15. López, P., Montes, M., Villasenor, L., Carrasco, A., & Martínez, J. (2012). A new document author representation for authorship attribution. In Mexican Conference on Pattern Recognition. Springer, 283-292
16. Qian. C, He. T, Zhang. R. (2017). Deep Learning Based Authorship Identification
17. Posadas-Durán, J. P., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Pinto, D., & Chanona-Hernández, L. (2017). Application of the distributed document representation in the authorship attribution task for small corpora. Soft Computing, 21(3), 627-639.
18. Potthast, M., Braun, S., Buz, T., Duffhauss. (2016). Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval. European Conference on Information Retrieval. Lecture Notes in Computer Science, 9626, 393-407, Springer
19. Plakias. S and Stamatatos, E. (2008). Tensor space models for authorship attribution. In Proc. of the 5th Hellenic Conference on Artificial Intelligence
20. Ramírez, J. M., Ruíz, M. C., & Somodevilla, M. J. (2014). Atribución de autoría combinando información léxico-sintáctica mediante representaciones holográficas reducidas. Res. Comput. Sci., 88, 103-113.

21. Queralt, S. (2014). Acerca de la prueba lingüística en atribución de autoría hoy. Revista de Llengua i Dret, (62).
22. Maciej. E, (2015). Does size matter? Authorship attribution, small samples, big problem. Digital Scholarship in the Humanities, 30(2), 167-182.
23. López-Monroy, A. P., Montes-y-Gómez, M., Villaseñor-Pineda, L., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2012). A new document author representation for authorship attribution. In Mexican Conference on Pattern Recognition Springer, pp. 283-292.
24. W. Anwar, I. Bajwa, S. Ramzan. (2019). Design and Implementation of a Machine Learning-Based Authorship Identification Model. Scientific Programming,1–14, DOI: 10.1155/2019/9431073
25. Alwajeeh, A., Al-Ayyoub, M., & Hmeidi, I. (2014). On authorship authentication of Arabic articles. In 2014 5th International Conference on Information and Communication Systems (ICICS), 1-6.
26. Gutiérrez, J. (2016). Enciclopedia de lingüística hispánica. Londres y New York: Routledge.
27. Flórez, R. y Fernández, R. (2008). Las redes neuronales artificiales, fundamentos teóricos y aplicaciones prácticas, Netbiblo, 33.
28. Solorio, T., Pillay, S., & Montes-y-Gómez, M. (2011). Authorship Identification with Modality Specific Meta Features. PAN, 1, 11.
29. Ledeneva, Y., García, R. (2017). Generación automática de resúmenes: retos, propuestas y experimentos", 1edición. México.
30. Houvardas, J., Stamatatos, E. (2006). N-Gram Feature Selection for Authorship Identification. Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications, Lecture Notes in Computer Science, 4183, 77–86.
31. Vel. O, Anderson. A, Corney. Malcolm, Mohay, G. (2001). Mining Email Content for Author Identification Forensics", ACM SIGMOD Record, 30, 55-64, DOI: 10.1145/604264.604272.
32. Koppel. M, Schler. J, Bonchek.E. (2007). Measuring Differentiability: Unmasking Pseudonymous Authors, Journal of Machine Learning Research, 8, 1261-1276, (2007).
33. Stamatatos. E. (2006). Authorship attribution based on feature set sub spacing ensembles, Int. J. Artif. Intell. Tools,15(5), 823–838.
34. Keselj. V, Peng. F, Cercone. N, Thomas. C. (2003). N-gram-based author profiles for authorship attribution. Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING 2003, pp. 255-264.
35. Benedetto. D, Caglioti. E, Loreto. V. (2002). Language trees and zipping, Physical Review Letters 88(4), 048702.
36. Koppel, M., Schler, J., Argamon, S. (2010). Authorship attribution in the wild. Language Resources and Evaluation, 45(1), 83–94,)
37. Stamatatos, E. (2007). Author identification using imbalanced and limited training texts. Proceedings International Workshop on Database and Expert Systems, DEXA, 237-241.
38. Teahan, W.J., Harper, D.J. (2003). Using compression-based language models for text categorization, Language Modeling for Information Retrieval, 141–165,
39. Peng, F., Schuurmans, D., Wang, S. (2004). Augmenting naive Bayes classifiers with statistical language models. Information Retrieval 7, 317-345
40. Burrows, J. Delta. (2002). A measure of stylistic difference and a guide to likely authorship. Literary and Linguistic Computing, 17(3),267-287.
41. Arun, R., Suresh, V., Veni, C. (2009). Stopword graphs and authorship attribution in text corpora, International Conference on Semantic Computing 2019, pp. 192-196,